ITEA Journal 2008; 29: 67–74

Best Practices for Developmental Testing of Modern, Complex Munitions

Capt Joshua Stults
780th Test Squadron, Eglin AFB, Florida

The growing cost and schedule constraints on government weapons development programs as well as their rising complexity increase the need for a decision theoretic-framework for product development. This framework must rely on insight gained from a variety of sources for test planning, test evaluation, and decision support. The best practices presented in this article for system-level developmental test planning and execution are collected from reported experience and criticism of industry and government product development programs. These practices and methodologies are applied in a coherent framework that allows a formal combination of the disparate sources of product knowledge available to decision makers in the early stages of development.

Key words: Bayes Theorem; best practices; complexity; external validation; knowledge-based acquisition; weapons systems.

This article illustrates a formal decision support framework for program managers and testers that embodies the ideas of knowledge-based acquisition and incorporates best practices identified from historical product development programs in the government and commercial sectors. Emphasis is on system-level developmental test and evaluation (DT&E) in support of risk reduction for production decisions. The framework consists of four basic steps: identify relevant system performance factors, use prior knowledge to evaluate system level outcomes, incorporate validated knowledge into product improvements and evaluate sufficiency of testing through external validation. The motivation for such a formal decision support framework is the growing complexity of modern weapon systems. While complexity is not easy to define or measure consistently, indicators of complexity are type and number of weapon sensors, multiple operational modes, multiple communications links, software for autonomous loitering or targeting, etc. These indicators have been shown to increase the cost of test and evaluation (T&E) despite the significant constraints currently being placed on weapons development funding (Fox et al. 2004).

The motivation for knowledge-based acquisition is to improve product development outcomes using "quantifiable and demonstrable knowledge to make go/no-go decisions" (GAO 2005). It is based on ensuring that the proper product knowledge is validated at critical decision points (DoD 2003). Central to this acquisition approach is the progression of the product through well-defined maturity levels, driven by validated product knowledge.

Three main product maturity levels have been identified through analysis of successful product development practices in industry. The product progresses through these levels based on specific events that demonstrate validated product knowledge rather than schedule driven milestones (GAO 2000). Heuristics learned from commercial and government product development programs can guide the planning of a knowledge validation (testing) program to successfully progress through the product maturity levels. Ideas such as "break it big early" are examples of these sorts of experience-based rules of thumb (GAO 2000).

In addition to informal rules of thumb, there are rigorous inference methods that can support knowledge validation and decision making even in the system development phase when sample sizes are too small for standard large sample size statistical methods to apply. For example, approaches based on Bayes theorem which incorporate prior knowledge in evaluating new knowledge as it arrives can ensure that product developers are making informed decisions even in the

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAR 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Best Practices for Developmental Testing of Modern, Complex Munitions	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **780th Test Squadron,Eglin AFB,FL,32542**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**8**	

face of few samples. Sequential Design of Experiments is another method that allows for smaller expected numbers of test events to achieve a given statistical power by using some sort of stopping rule (Cohen and Rolph 1998).

The product maturity paradigm, experience-based heuristics, formal inference and design of experiments methods can be tied together into a coherent decision support framework by a high-fidelity system performance model as suggested in (Cohen and Rolph 1998). System performance models provide a repository for the product knowledge gained as the system matures, so that successive testing can be planned based on validated knowledge. They can support a constructive approach to testing that leverages knowledge discovery from the early phases of product maturity for more efficient system level DT&E. Likewise, as has been previously suggested, the knowledge gained from DT&E to develop and validate the system performance model should be used for efficient operational test and evaluation (OT&E) planning (Cohen and Rolph 1998).

A recurring criticism of Department of Defense product development is that programs proceed without the right kind of knowledge gained from test efforts. When this happens cost, schedule, and performance problems often result (GAO 2003). As has been observed, "It is possible to conduct a test or simulation that does not contribute worthwhile information" (GAO 2003). By focusing on knowledge validation and knowledge driven product maturity rather than specific test schedules or events, we hope to avoid this waste of effort and ensure that all planned test events validate the right knowledge at the right level of product maturity.

Product maturity levels

Three levels of product maturity identified in (GAO 2000) are:

1. Technologies and subsystems work individually;
2. Components and subsystems work together as a system in a controlled setting;
3. Components and subsystems work together as a system in a realistic setting.

This article will focus on the second and third levels of product maturity which correspond to system-level DT&E. Oftentimes because the number of system-level tests during the DT&E phase of weapon development is not large enough for statistical significance in the classical frequentist sense, these tests are relegated to "demonstration" status. When incorporated into a Bayesian inference framework, these tests can support a meaningful estimate of parameters important to programmatic decisions from the first test event. In addition, the marginal value (reduction in risk) of additional testing can begin to be compared to the marginal cost of that testing. This comparison is critical to allowing for a decision theoretic approach to answering the question of how much testing is enough (Cohen and Rolph 1998).

Knowledge validated by testing drives the progress of a product through the stages of development. Incorporating the knowledge gained from each phase of testing and development can guide the test plan to be more efficient than starting from assumed ignorance at each stage. Assuming ignorance is conservative as far as technical risk goes, it drives larger and less efficient test plans than if prior knowledge is incorporated into the planning effort.

Historically based heuristics for test planning and product development

A very disciplined approach to maturing a product is required to avoid costly rework late in product development. The three critical factors that underlie this disciplined approach ensure that:

1. Validation is event based rather than schedule based;
2. The quality of the knowledge validated in each event is not sacrificed;
3. The knowledge validated in each event is used to improve the product (GAO 2000).

One of the most important heuristics identified from successful commercial product development efforts is known as "break it big early", or "move discovery to the left" (GAO 2000). This means that challenging validation events are planned early to expose areas of weaknesses in the new design.

Rigorous subsystem verification has been identified as one of the means to reduce the burden of discovery on the later system level test events. This is a way to ensure that the quality of knowledge gained from test events does not suffer due to immature test articles. Aggressive development schedules can often result in an undue burden of discovery on system-level flight testing. Experience in the Theater High Altitude Air Defense (THAAD) program illustrated that shortcomings in component and subsystem validation lead to very expensive failures in the flight test program (GAO 2000). Sacrifices were made in the first two stages of product maturity to keep system level flight testing on schedule. The problems experienced by THAAD were not that tests failed or discoveries occurred, which is the very purpose of testing. In fact, it has been pointed out that "...bad things happen in test and that those bad things are valid results just as successes are" (DOT&E 2007). The object is to find those bad things early in component level and

subsystem integration testing, so that the discoveries during more expensive full-up system level testing are small and affordably corrected.

Also in line with the "break it big early" philosophy is to test at factor levels that give the most variation in system performance. System response in most real systems is nonlinear, so the factor level matters. The most knowledge can be gained from a limited number of test events by testing at the most stressing factor levels.

In keeping with the third element of disciplined product development, information gained from initial test events must be incorporated into improving the product. Using knowledge to mature the product and getting the right knowledge to decision makers is the focus rather than sacrificing the quality of test events to maintain schedule goals. The DarkStar Unmanned Aerial Vehicle program experienced significant flight test failures and was eventually terminated due to problems that surfaced during initial flight testing which were not addressed and fixed before subsequent testing continued (GAO 2000). The point here is not that flight test failures cause program termination, but that sacrificing knowledge validation and product improvement based on validated system knowledge to maintain schedule is counterproductive.

If these heuristics are applied to the first two levels of product maturity, then the burden of discovery on system-level DT&E will be reduced (GAO 2005). This allows more operational realism to be incorporated into DT&E, thus improving the quality of knowledge gained from these test events.

The Stand-off Land Attack Missile – Expanded Response (SLAM-ER) system experienced failures during OT&E that were masked in earlier testing because of unrealistic DT&E test conditions and immature test articles (GAO 2000). This shows how the heuristics identified can complement each other, mature test articles support more operational realism in DT&E which in-turn supports "moving discovery to the left."

To summarize the above discussion, here is a collection of some of the experience-based rules of thumb:

- *Break it big early, move discovery to the left*
  - Rigorous subsystem verification and integration minimizes discovery burden on the final, most expensive system-level development effort;
  - Test difficult technology or design features early;
  - Test at factor levels that give the most variation in system performance: System response in most real systems is nonlinear, the level matters.
- *Focus on getting necessary knowledge to decision makers rather than specific events, techniques, or schedules*
  - Incorporate information from early test events to improve the product before proceeding to future test events;
  - Do not curtail early testing to stay on schedule;
  - Do not sacrifice test-item fidelity to stay on schedule: Unrealistic system level test events lower the amount of useful information gained from those events.

Importance of system performance models

Incorporating knowledge gained from disciplined component and subsystem validation into a high-fidelity system performance model informs decision makers about development and production risk. This can also lead to more efficient test planning and analysis. The system performance model tracks the system through the product maturity levels. As product knowledge is validated in each level, that knowledge is incorporated into the model. The model provides a means for the heuristics identified in Section 3 to be rigorously applied. It allows the test planner to answer the questions like:

- Where can I expect the most variation?
- What level of product maturity is the modeled performance based on?
- What discoveries have been made, and has that knowledge been incorporated into the product (and its model)?

The test planner can make basic decisions about influential factors and their likely critical levels before design details of the actual test article are finalized. In other words, "one can design an effective test for a system without understanding precisely how a system behaves" (Cohen and Rolph 1998). This allows testing for the later levels of product maturity to be based on knowledge gained during the initial levels. *Figure 1* illustrates the progression of model maturity. Initially, the insight for test planning comes from physics-based simulation and other analysis tools. As the product matures and component and integration testing data become available these can be used for test planning and decision making. The fast running engineering models are based on the more fundamental information in the detailed physical models. Component performance and integration testing data are incorporated as they become available.

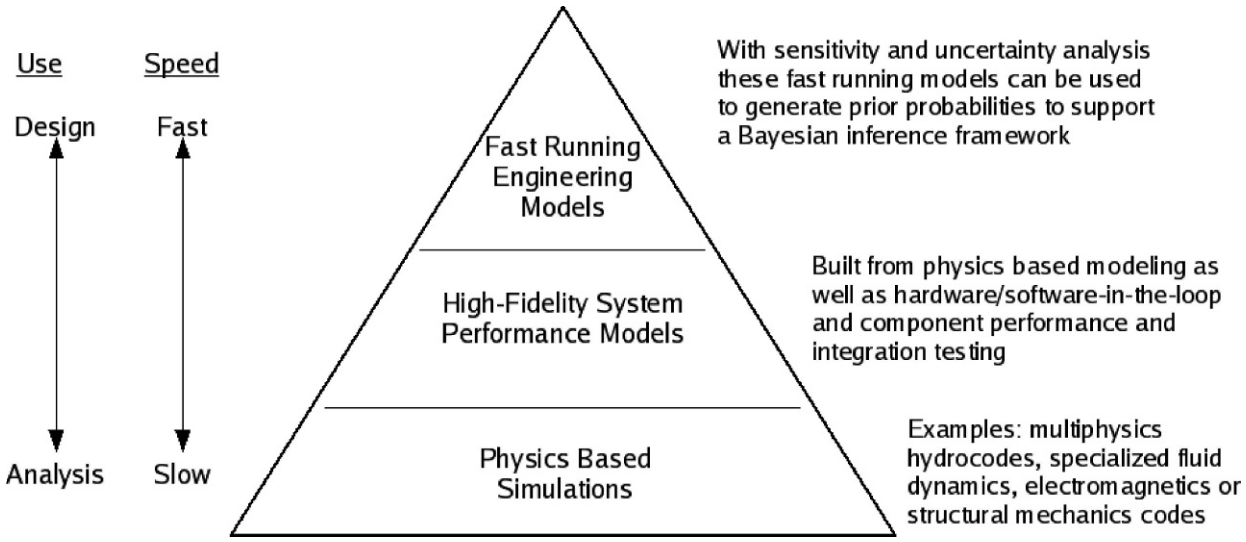

Figure 1. Modeling hierarchy

Incorporating prior knowledge

Knowledge captured in the system performance model (based on component level testing and system design analysis) can be used to generate prior probabilities in performance metrics of interest. These prior probabilities, or degrees of belief, are useful for a Bayesian inference method.

The Bayesian approach has advantages over approaches which do not adjust their prior probabilities based on experience (Robbins 1964). It is desirable because it gives an optimal prediction: given the hypothesis prior probabilities, any other prediction will be correct less often (Russell and Norvig 1995). Bayes Theorem is shown in Equation 1.

$$P(H_j|E_i,I) = \frac{P(E_i|H_j,I)P(H_j|I)}{P(E_i|I)} \qquad (1)$$

Where the posterior, or final, probability of the hypothesis, H_j , being true given the new data, E_i, and the background information, I is updated by the likelihood, $P(E_i|H_j, I)$, and the prior or initial probability, $P(H_j|I)$. Beliefs about the system under test are updated by new information gained from each test event.

A common criticism of the Bayesian approach is that there is subjectivity in choosing the prior probabilities. This is true, but the benefit is that an explicit exposition of the assumptions underlying the test planning and analysis has been made, which is often not the case for other test planning approaches. In addition, the dependence of the result on the prior probability decreases as the sample size increases. In the large sample size limit, for certain model assumptions the Bayesian approach matches the more standard frequentist result (D'Agostini 2003).

High level test planning for weapon development programs tends to focus on the number of end-to-end flight tests because this is a significant contribution to overall test program cost and schedule. Performing enough end-to-end testing to build confidence intervals based on large sample-size theories is cost and schedule prohibitive, so the end-to-end testing is many times relegated to a demonstration only status. If the system level test events are merely demonstration, there is little rigorous or quantifiable connection between those small samples and knowledge gained to support decision criteria.

Since there is no quantifiable connection the argument is often put forth that a sample of 1 is as good as 1 + m, where m is some number small enough that large sample theories still do not apply with sufficient power. This argument is fallacious because large sample theory is not meant to measure the difference in marginal information gained between two small samples. It does not follow that there is no difference in value to the decision maker because large sample theories cannot measure that difference.

A Bayesian approach incorporates assumptions and prior knowledge about the system under test in a formal way so that information gained beginning with the first test event improves the certainty of the knowledge about the system in a quantifiable manner. Some estimation of the marginal value of n and $n + 1$ samples can be evaluated even though n is far too small for frequentist statistical approaches to apply. There is no free lunch here. With very small n the inferences supported by a Bayesian approach will be quite sensitive to the priors; however, that sensitivity information can be provided to decision makers so

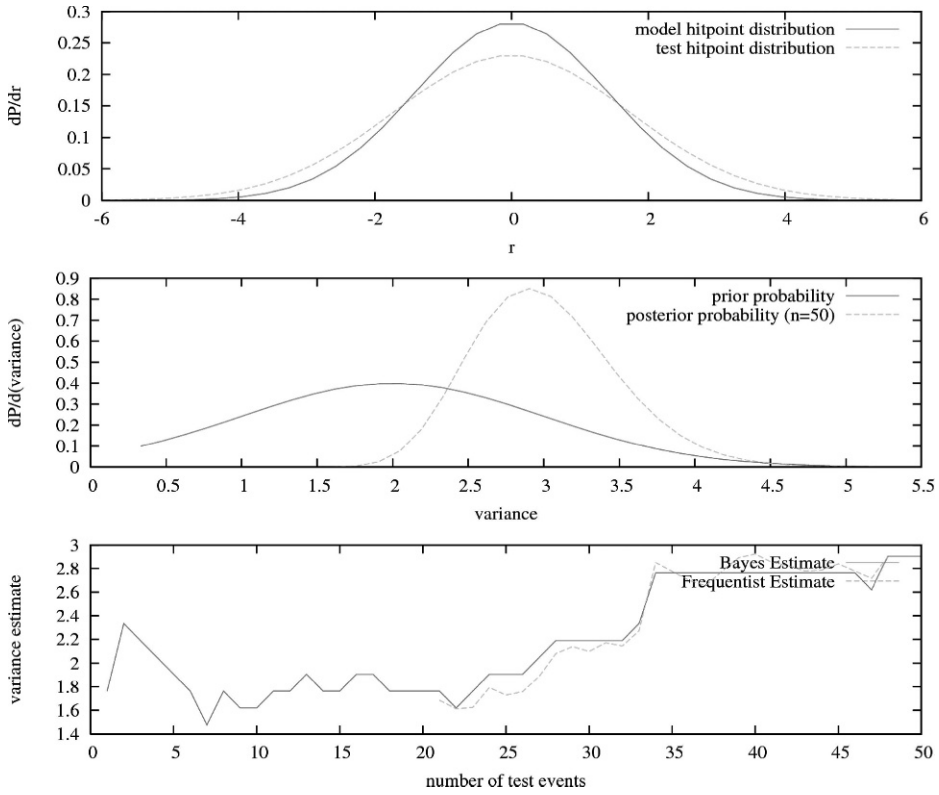

Figure 2. Estimating variance in hit-point distribution

that they understand what increasing n will mean in terms of reduced risk.

Hit-point distribution

This section presents an example of the Bayesian approach evaluating hit-point distributions for a munition with some type of smart terminal guidance based on a multimode seeker and target recognition algorithms. The seeker component level testing and closed-loop guidance and control simulation can provide a probability density for the hit-point in the plane normal to the weapon's attack vector. This information provides a prior probability for evaluating the hit-point from the very first end-to-end flight test. For smaller, smarter munitions this hit-point becomes increasingly important. Great variations in system effectiveness (i.e., killing the target) might be expected for small variations in hit-point.

Figure 2 illustrates using the Bayesian approach to estimate the variance in hit-point distribution. The model predicts a radial distribution of hit-points with a variance of two, while the actual performance is drawn from a distribution with variance of three. The variance in this example is our hypothesis, and the prior probabilities (see Equation 1) for the hypothesis could be generated from sensitivity and uncertainty analysis of the model. The actual form for the prior is not critical as long as there is some finite probability assigned to the true answer (Russell and Norvig 1995).

The lowest graph in *Figure 2* shows the maximum probability estimate of the Bayes method and compares it to the standard frequentist result (for $n > 20$). Rather than integrate over the continuous hypothesis space (variance in this case), a discrete set of hypotheses is evaluated. This is why the Bayesian estimate in *Figure 2* jumps discontinuously between levels. The method allows significant insight into the problem while the sample size is still small compared with more standard estimation methods.

Model output for prior probabilities

Suppose the output of an uncertainty analysis for a simple fast-running model can be given by Equation 2,

$$y = \beta_0 + e_0 + (\beta_1 + e_1)x \qquad (2)$$

where $\beta_0 = 1$, $\beta_1 = 3$, and e_0 , e_1 are normally distributed errors with zero mean and 0.25 standard deviation. The variation simulated here by e_0 , e_1 can be generated by sensitivity and uncertainty analysis in a fast running engineering model. The prior distributions for the model parameters can be estimated by holding the other parameters constant at their expected value and treating each data point as a measurement of the parameter of interest.

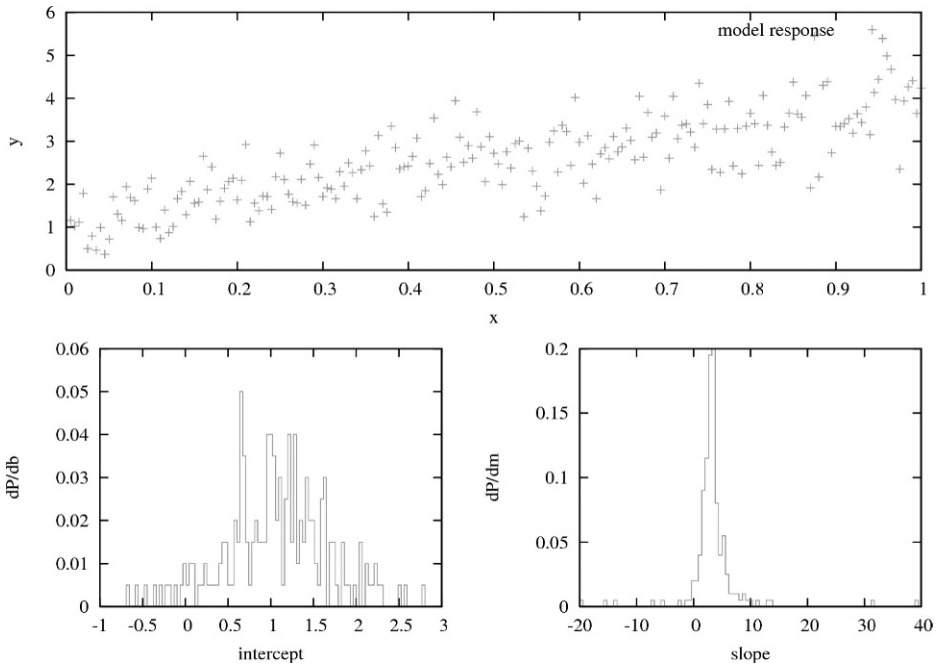

Figure 3. Estimation of prior probability from model output

Figure 3 shows the probability distributions for the slope and the intercept of the model's output following this method. These prior probabilities can be used to guide test planning by identifying where variation or uncertainty is greatest, which leads naturally to where testing will be most profitably executed. The best practice heuristics previously discussed become more than just good rules of thumb when informed by a Bayesian planning and analysis framework. This framework provides insight into where the variation in system performance can be expected, because it explicitly incorporates the prior knowledge from component-level testing residing in the system performance model.

Sequential design of experiments

The basic idea of sequential design of experiments is to test progressively from the outside of the parameter space, capturing linear effects, towards the inside of the parameter space, capturing higher-order interaction effects if needed (Curry and Lee 2007). A comprehensive review of the field is given in (Lai 2001). At each level, the predictive power of the effects measured so far is evaluated and a decision is made about whether additional testing is required.

For example, perhaps the product development team has identified some significant factors for a notional munition with terminal phase guidance and in-flight communication as follows: target aspect (TA), target speed (TS), target movement duty cycle (TMDC), impact angle (IA), engagement mode (EM), and target type (TT). Factors such as noise environment or weather are generally uncontrollable by the testers, but it is worthwhile to note their significance and then record their levels during test events so their influence on performance can be quantified (Cohen and Rolph 1998).

An initial experimental design will attempt to measure the linear or "main" effects. For the six controllable factors identified above, a seven-parameter model results, requiring seven tests at the minimum to make point estimates of the parameters (shown in Equation 3). Two additional tests are added to the design so that some estimate of the process variability can be made, and a final confirmation test is added to evaluate the sufficiency of the linear model.

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \qquad (3)$$

Given ten test events and minimum and maximum levels for each of the factors, a constrained optimization method can be applied to find the combination of factor levels across the tests that gives the lowest factor correlation. This is known as a d-optimal test design since it maximizes the determinant of the factor correlation matrix (Curry and Lee 2007).

One method of reaching an approximate optimum is simulated annealing (exactly orthogonal test series exist only at multiples of four tests). It is a heuristic optimization method that combines both divide-and-conquer and iterative improvement strategies (Kirkpatrick and Gelatt 1983). The method starts with a feasible set of factor levels for the test series and then swaps factor levels and evaluates if this improves or

Table 1. Approximately d-optimal test design

Test	TA	TS	TMDC	IA	TT	EM
1	360	20	0.1	15	1	1
2	360	20	0.9	75	−1	−1
3	180	4	0.9	15	−1	1
4	360	20	0.9	15	−1	−1
5	180	4	0.9	15	1	1
6	180	20	0.9	75	1	1
7	180	4	0.9	15	1	−1
8	180	20	0.9	75	−1	1
9	360	4	0.1	75	1	1
10	180	4	0.1	75	−1	−1

TA, target aspect; TS, target speed; TMDC, target movement duty cycle; IA, impact angle; TT, target type; EM, engagement mode.

degrades the orthogonality of the tests. If the change improves the orthogonality, it is accepted with probability, $P = 1$. If the change degrades the orthogonality, it is accepted with probability relation shown in Equation 4.

$$P = e^{d_1 - d_0 / T} \quad (4)$$

Where d_1 is the determinant of the correlation matrix (a measure of orthogonality or "goodness") and T is the temperature, a parameter that is gradually reduced during the optimization. This allows the process to avoid being trapped by local minima because it accepts moves which are "bad" according to the difference $d_1 - d_0$ and the cooling schedule in T. As cooling progresses the algorithm accepts "bad" moves with less and less probability.

A test series developed by the simulated annealing method is shown in *Table 1*. The correlation of factors across the test events for this design is shown in *Table 2*.

An exactly orthogonal series would have no nonzero off-diagonal terms in the correlation matrix. The goal of the optimization is to make these terms approximately zero. The advantage of using an optimization technique like simulated annealing is that constraints on the test design can easily be added and optimization can proceed exactly as before, only within the reduced set of feasible designs. For example, the factors describing an important operationally representative scenario can be constrained to occur a given number of times.

Importance of external validation

In a test program that relies heavily on modeling and simulation, it is critical to guard against over-fitting the model. The basic algorithm to avoid such over-fitting is known as "model-test-model-test" (Cohen and Rolph 1998). The final validation tests are outside the scenarios which were used for parameter tuning. Sequential design of experiments naturally provides the framework for such an approach. The stopping rule in a standard sequential design depends on evaluating the predictive power of the simple empirical model using the final additional test.

When a high-fidelity system performance model is available the stopping rule should be modified to depend on an external validation of the system performance model as well as the more standard stopping rule. The initial tests used to develop the simple linear empirical model can also be used for parameter tuning of the high-fidelity model and the final test serves as an external validation of the high-fidelity model as well.

Conclusions

High-fidelity system performance models along with full-up system level test events incorporated into a formal inference framework provide rigorous support to decision makers in developing and acquiring modern weapon systems of ever-increasing complexity. The proposed framework for knowledge-based test planning and execution consists of four basic steps:

1. Identify significant factors and levels based on a high-fidelity system performance model;
2. Use the model for prior distributions (context, background knowledge) with which to analyze full-up system level test outcomes;
3. Incorporate discoveries into product improvements and improved performance model;
4. Evaluate sufficiency of testing based on predictive power of high-fidelity system performance model, i.e., model-test-model-test.

Table 2. Factor cross-correlation matrix

	TA	TS	TMDC	IA	TT	EM
TA	1	0	0	0	0.2	0
TS	0	1	−0.16667	0.16667	0	0.102062
TMDC	0	−0.16667	1	−0.16667	0	−0.102062
IA	0	0.16667	−0.16667	1	0	0.102062
TT	0.2	0	0	0	1	0
EM	0	0.102062	−0.102062	0.102062	0	1

TA, target aspect; TS, target speed; TMDC, target movement duty cycle; IA, impact angle; TT, target type; EM, engagement mode.

The exact mechanics of the approach presented in this article are not critical. Any integrated method that gives some measure of the marginal value of system-level test events when sample sizes are small can provide useful support to decision makers. This support will begin to allow hard risk management decisions about how much testing is sufficient to be made in a more decision-theoretic framework.

The critical aspect of the approach is the knowledge warehouse known as the system performance model. The knowledge it contains at the same time informs decision makers and test planners, and provides a repository of validated knowledge from test conductors. The execution of a knowledge-based test program supports decision makers with solid information about test sufficiency and risk. Through improvements incorporated into the product and its model, it ensures that decisions made about the system are based on the highest quality of information available. ❑

CAPT JOSHUA A. STULTS is the deputy live fire agent for conventional munitions, Air Force Live Fire Office, 780th Test Squadron, Eglin AFB, FL. Prior to his current duties supporting U.S. Air Force weapons programs in planning and executing live fire test and evaluation, he was a test engineer for the 780th supporting weapons development flight test. He holds a master of science degree in aeronautical engineering from the Air Force Institute of Technology, Dayton, Ohio; and a bachelor of science degree in aeronautical engineering from the U.S. Air Force Academy, Colorado Springs, Colorado. E-mail: joshua.stults@us.af.mil

References

Cohen, M. L. and Rolph, J. E. (eds.) 1998. "Statistics, Testing and Defense Acquisition: New Approaches and Methodological Improvements." *Washington, D.C.: National Academy Press*.

Curry, T. F. and Lee, S. J. 2007. "Using Sequential-Designed Experimentation to Minimize the Number of Research and Development Tests." *The ITEA Journal of Test and Evaluation,* Volume 28-2, pp. 41–47.

D'Agostini, G. 2003. "Bayesian Inference in Processing Experimental Data: Principles and Basic Applications." *Rept. Prog. Phy*. 66, pp. 1,383–1,420.

DoD, (Undersecretary of Defense for Acquisition, T. and Logistics). 2003. "Operation of the Defense Acquisition System, No. 5000.2." In: *Department of Defense Instructions.* May 2003, pp. 1–50.

DOT&E. 2007. *"Lessons Learned from Live Fire Testing: Insights Into Designing, Testing, and Operating U.S. Air, Land, and Sea Combat Systems for Improved Survivability and Lethality."* O'Bryon, J. F., (ed). Washington, D.C.: Secretary of Defense, Operational Test and Evaluation Directorate (DOT&E), Live Fire Test and Evaluation, pp. 3–15.

Fox, B., Boito, M., Graser, J. C. and Younossai, O. 2004. "Test and Evaluation Trends and Costs for Aircraft and Guided Weapons, Tech. Rep. MG-109." Arlington, Virginia: RAND Corporation.

GAO (General Accounting Office). 2000. "Best Practices: A More Constructive Test Approach is Key to Better Weapon System Outcomes, Tech. Rep. GAO/NSIAD-00-199." Washington, D.C.: U.S. General Accounting Office. Available online at http://www.gao.gov. Accessed January 18, 2008.

GAO (General Accounting Office). 2003. "Defense Acquisitions: Assessment of Major Weapon Programs, Tech. Rep. GAO-03-476." Washington, D.C.: U.S. General Accounting Office. Available online at http://www.gao.gov. Accessed January 18, 2008.

GAO (General Accounting Office). 2005. "Best Practices: Better Support of Weapon System Program Managers Needed to Improve Outcomes, Tech. Rep. GAO-06-110." Washington, D.C.: U.S. General Accounting Office. Available online at http://www.gao.gov. Accessed January 18, 2008.

Kirkpatrick, S. and Gelatt, M. V. 1983. "Optimization by Simulated Annealing." *Science,* Volume 220, No. 4598, pp. 671–781.

Lai, T. L. 2001. "Sequential Analysis: Some Classical Problems and New Challenges." *Statistica Sinica,* Volume 11, pp. 303–408.

Robbins, H. 1964. "The Empirical Bayes Approach to Statistical Decision Problems." *The Annals of Mathematical Statistics,* Volume 35, No. 1, 1964, p. 1.

Russell, S. and Norvig, P. 1995. "Artificial Intelligence: A Modern Approach." 2nd ed. Upper Saddle River, NJ: Prentice Hall. p. 1132.

Acknowledgments

The author thanks Maj David Winebrener for the lively discussions on test planning and lessons learned that led to this research effort.